Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Short-term PM10 exposure and hospital admissions for respiratory illness in children under five in Rio de Janeiro: a distributed lag study with negative-control evaluation of fifteen estimators

Ankita Mishra

ankita.mishraphd2020@gmail.com

Department of Computer Science and Engineering Banasthali Vidyapith Rajasthan, India

Yogesh Kumar Gupta

Department of Computer Science and Engineering Banasthali Vidyapith Rajasthan, India

Research Article

Keywords: PM10, distributed lag model, pediatric respiratory admissions, negative control exposure, model specification, Rio de Janeiro

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11099089/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Short-term PM10 exposure and hospital admissions for respiratory illness in children under five in Rio de Janeiro: a distributed lag study with negative-control evaluation of fifteen estimators

[1*]Ankita Mishra, [2]Yogesh Kumar Gupta
[1,2]Department of Computer Science and Engineering Banasthali Vidyapith
Rajasthan, India
ankita.mishraphd2020@gmail.com[1*], gyogesh@banasthali.in[2]

[*]**Corresponding Author:** Ankita Mishra (ankita.mishraphd2020@gmail.com)

Abstract

Daily time-series studies of air pollution and health rest on modeling choices that fit statistics cannot validate, since a bias shared by every candidate model is invisible to criteria computed on the same data. This study linked daily hospital admissions for respiratory illness among children under five in Rio de Janeiro to daily PM10 from eight municipal monitoring stations over 2011 to 2019, fitting a negative binomial distributed lag model across a 14-day window. Every estimator was refitted to exposure series shifted by one and two whole years, which preserves seasonality while breaking day-to-day correspondence, so any surviving association is not an exposure effect. Across 37,137 admissions, a 10 μg/m³ increase in PM10 was associated with a 3.75% rise in admissions (95% CI 0.63 to 6.97), of which same-day exposure recovered only 1.67%, and an estimated 449 admissions (80 to 795) were attributable to concentrations above the World Health Organization 2021 daily guideline. Eleven of fifteen estimators passed the control and four neural estimators failed. Seasonal adjustment at four degrees of freedom per year returned a near-null estimate whose shifted controls reached 6.53 percentage points, revealing broken identification rather than absence of an effect.

Keywords: PM10; distributed lag model; pediatric respiratory admissions; negative control exposure; model specification; Rio de Janeiro

1. Introduction

Children under five are the age group most affected by ambient particulate pollution. Their airways are still developing, they breathe more air relative to body weight than adults, and respiratory infection is already the leading cause of hospitalization in this group in Brazil (Freitas and Donalisio 2016). Hospital admissions are a useful outcome for quantifying that burden, since they are recorded routinely, they are severe enough to matter for policy, and they are counted daily, which makes them suitable for the time-series designs that dominate this field.

The evidence base is substantial but the pooled estimates disagree. Meta-analyses of short-term particulate exposure and pediatric respiratory outcomes report anywhere from 0.4% per 10 μg/m³ of PM10 for pneumonia admissions (Yee et al. 2021) to 1.75% for asthma admissions (Ding et al. 2015), with between-study heterogeneity reaching an I^2 of 95.7% (Lim et al. 2016). Some of that spread reflects genuine differences between populations. Much of it reflects differences in how the studies were specified, in the lag window used, the way season was adjusted for, and the estimator fitted. Those are analyst choices, not properties of the data.

Two of those choices are worth separating. The first is the lag window. Pollution does not act only on the day it is inhaled; a child exposed on one day may be admitted several days later. Distributed lag non-linear

models were developed to represent this, writing the association as a surface over concentration and elapsed time (Gasparrini 2014; Gasparrini and Armstrong 2013; Gasparrini et al. 2010). Yet many applied studies, including several Brazilian analyses using the same hospital system used here, continue to work with same-day concentration or a moving average over five or six days (Nascimento et al. 2025; Negrisoli and Nascimento 2013; Tuan et al. 2015). Whether that costs anything, and how much, is an empirical question that a longer window can answer directly.

The second choice is seasonal adjustment, and it is the harder one. Particulate concentration is seasonal. Childhood respiratory illness is seasonal, driven largely by respiratory syncytial virus, whose timing in Brazil is well characterized and regionally specific (Freitas and Donalisio 2016; Julian et al. 2026; Pereira et al. 2023). An analysis that adjusts for season too loosely will attribute viral seasonality to pollution; one that adjusts too aggressively will remove the exposure signal along with the season. Practice has converged on roughly seven degrees of freedom per year, but this is a convention rather than a quantity derived from the data, and Peng and colleagues showed two decades ago that estimates move materially with it (Dominici et al. 2005; Peng et al. 2006). A recent Brazilian multi-municipality study reported that the selected degrees of freedom varied considerably across the municipalities it analyzed (Gouveia et al. 2017).

The usual defense of these choices is an information criterion. That defense has a specific and underappreciated limit, since AIC, deviance and held-out likelihood are all computed from the same data used to fit the model, so they can rank two candidate specifications but cannot detect a bias that both share. If residual seasonal confounding inflates every candidate equally, no fit statistic will say so.

Negative controls provide the external check that fit statistics cannot. Lipsitch et al. (2010) formalized the idea for observational epidemiology, and Flanders et al. (2011, 2017) adapted it to air pollution time series by using future pollution as a control exposure, on the reasoning that tomorrow's concentration cannot have caused yesterday's admission. The framework has been reviewed and extended since (Hu et al. 2023; Shi et al. 2020). Despite this, negative controls remain rare in applied air pollution studies, and are almost never reported alongside the headline effect estimate.

The gap matters more now than it did a decade ago. Machine learning and neural network methods are being proposed for exposure-response estimation in this setting, on the argument that they impose fewer assumptions about functional form (Abedian Aval et al. 2026; Agbehadji and Obagbuwa 2025; Sun et al. 2023). Published evaluations of these methods report goodness of fit and sometimes held-out prediction. No published example was found in this application that reported whether the proposed model returns a null result when the exposure carries no signal. A flexible estimator that fits well and predicts well may still be recovering season rather than pollution, and neither of the reported diagnostics would distinguish the two cases.

This study addresses that gap using nine years of daily data from Rio de Janeiro, a city of six million with eight municipal air quality monitors and a public hospital system whose records are openly downloadable. The analysis estimates the association between PM10 and respiratory admissions in children under five, and then subjects every estimator fitted to a negative control constructed by shifting the exposure series by whole years. A whole-year shift preserves the seasonal cycle, the autocorrelation and the marginal distribution of the exposure series, breaking only its day-to-day correspondence with admissions. Any association that survives it is not an exposure effect, and because season is preserved, the control is targeted at residual seasonal confounding specifically.

This paper makes five contributions.

First, it provides a distributed lag estimate of the association between PM10 and under-five respiratory hospitalization in Rio de Janeiro over 2011 to 2019, based on 37,137 admissions across 3,287 days, with an attributable burden quantified against the World Health Organization 2021 daily guideline. Both data sources are public and require no registration or institutional agreement, so the analysis is reproducible end to end.

Second, it quantifies what a same-day exposure metric costs. Fitting the same model with same-day concentration alone recovered 1.67% per 10 μg/m³ against 3.75% over a fourteen-day window, so more than half the cumulative association fell outside the exposure day. The per-lag estimates showed why, in that no single lag dominated and the effect accumulated in small daily increments across the fortnight.

Third, it reports what is, to the authors' knowledge, the first systematic negative-control comparison of exposure-response estimators in this application. Fifteen estimators were fitted to one dataset with one confounder specification, so that any difference between them arose from the exposure specification alone. The set spanned single-lag, moving-average, polynomial, spline, unconstrained and ridge-penalized distributed lag models, two tree ensembles, and four variants of three neural architectures. Each is reported with four shifted controls and a pass criterion fixed in advance.

Fourth, it shows that the same control can select the amount of seasonal adjustment, replacing a convention with an external criterion. At four degrees of freedom per year the model returned an estimate near zero, which read alone would be reported as a null finding; its shifted controls returned −6.53 pp, revealing that the specification recovered large associations from exposure carrying no information. No fit statistic computed on those data disclosed this.

Fifth, it documents that four neural exposure-response estimators fail the negative control that eleven classical and tree-based estimators pass on identical data. Simulation shows that the cross-fitted estimator is approximately unbiased but has a standard deviation roughly three times that of the classical model, so its failure reflects variance rather than systematic bias. Validation-based early stopping is identified as one contributor to the failure of the free-surface variant, and the residual cause is reported as unresolved.

The remainder of the paper is organized as follows. Section 2 reviews the relevant literature across pediatric particulate epidemiology, Brazilian time-series studies, distributed lag models and seasonal confounding, negative controls, machine learning applications, and viral seasonality. Section 3 describes the data and their processing. Section 4 sets out the model, the negative control and the estimator bank. Section 5 reports results, and Section 6 discusses them.

2. Related work

2.1 Particulate matter and pediatric respiratory morbidity

The association between short-term particulate exposure and respiratory morbidity in children has been examined repeatedly, though pooled estimates vary with outcome definition, pollutant and design. A meta-analysis of 21 studies of pneumonia admissions and emergency visits reported a 1.0% increase (95% CI 0.5 to 1.5) per 10 μg/m³ of PM2.5 and 0.4% (0.2 to 0.6) per 10 μg/m³ of PM10 (Yee et al. 2021). For pediatric asthma specifically, a meta-analysis of 26 time-series and case-crossover studies found a relative risk of 1.048 (1.028 to 1.067) per 10 μg/m³ of PM2.5, with regional variation and effects in Asia lower than in North America or Europe (Lim et al. 2016). An earlier meta-analysis restricted to children's asthma

admissions reported 1.75% per 10 μg/m³ for PM10 and 3.45% for PM2.5, and found case-crossover designs producing larger estimates than time series (Ding et al. 2015). Weinmayr et al. (2010), synthesizing 36 studies of children with asthma or asthma-like symptoms, reported a significant association of PM10 with asthma symptoms and noted stronger associations for studies using longer lags. The lag results reported here bear on that observation directly.

These pooled figures span roughly 0.4% to 3.5% per 10 μg/m³ depending on outcome and pollutant. The heterogeneity across meta-analyses, which reached an I^2 of 95.7% in one (Lim et al. 2016), reflects differences in design as much as in populations, which is part of the motivation for comparing estimators on a single dataset rather than across studies.

2.2 Brazilian time-series studies using the public hospital system

Brazil's Sistema de Informações Hospitalares (SIH-SUS) has supported a substantial body of time-series work, which makes it possible to place the present estimates in a directly comparable setting.

Gouveia and Fletcher (2000) analyzed daily pediatric respiratory admissions in São Paulo and reported increases of about 9% for PM10 across the 10th to 90th percentile contrast, with pneumonia showing larger effects than respiratory disease overall. Later studies covered smaller and medium-sized cities, among them pneumonia admissions in Sorocaba, using lags of zero to five days (Negrisoli and Nascimento 2013); pneumonia in São José dos Campos (Tuan et al. 2015); and acute respiratory admissions in the conurbated cities of Taubaté, Tremembé and Pindamonhangaba, analyzed with negative binomial regression, which reported significant PM10 associations at all lags except lag 4 (Nascimento et al. 2025). A multi-municipality analysis of the São Paulo metropolitan region fitted semiparametric Poisson models per municipality and combined them by meta-analysis, noting that degrees of freedom for seasonal control varied considerably between municipalities according to AIC and autocorrelation criteria (Gouveia et al. 2017). Estimates of admissions avoidable under revised Brazilian air quality standards have been produced for four southeastern capitals including Rio de Janeiro (de Oliveira Fernandes et al. 2020).

Two features of this body of work are relevant here. Several studies used lag windows of five or six days, or same-day and short moving-average exposure metrics (Nascimento et al. 2025; Negrisoli and Nascimento 2013; Tuan et al. 2015). Several studies also worked with small daily counts, the Paraíba Valley study averaging 0.88 admissions per day (Nascimento et al. 2025). As Table 1 shows, the series analyzed here averages 11.3 admissions per day over nine years, which supports a more flexible lag structure than most of this literature could accommodate.

2.3 Distributed lag models and seasonal confounding

The cross-basis formulation of Gasparrini et al. (2010) represents the exposure-lag-response surface as a tensor product of a basis over exposure and a basis over lag, unifying earlier moving-average and polynomial distributed lag approaches within one framework. The method was extended beyond time series to general exposure histories (Gasparrini 2014), equipped with procedures for reducing fitted surfaces to one-dimensional summaries (Gasparrini and Armstrong 2013), and implemented in the R package dlnm (Gasparrini 2011). Gasparrini and Leone (2014) subsequently defined attributable risk measures within the distributed lag framework, distinguishing forward and backward formulations. The backward formulation is used here, as given in Equation (6).

Methodological development continues, including penalized and multi-pollutant variants. Bhaskaran and colleagues provided the standard practical account of time-series regression in this setting, covering seasonal control, time-varying confounding and lag specification (Bhaskaran et al. 2013).

Peng et al. (2006) examined how estimates of the pollution-mortality association depend on the degrees of freedom used for seasonal adjustment and on model selection strategy, and showed that the choice materially affects results. Reanalyses of the National Morbidity, Mortality and Air Pollution Study documented comparable sensitivity across 90 US cities (Dominici et al. 2005). Reported practice converged on roughly six to eight degrees of freedom per year, which now functions as a convention rather than a data-driven quantity. The Brazilian metropolitan study cited above explicitly noted large variation in the degrees of freedom selected across municipalities (Gouveia et al. 2017).

The difficulty is that information criteria compare candidates on the fitting data. The results in Table 5 and Figure 6 include a case in which the convention succeeds while a nearby specification fails, with the failure invisible in the model's own output.

2.4 Negative controls

Lipsitch et al. (2010) formalized negative controls for detecting confounding and bias in observational studies, distinguishing negative control exposures from negative control outcomes. Flanders et al. (2011) adapted the idea to air pollution time series, using future pollution as a negative control exposure on the reasoning that tomorrow's ozone cannot have caused yesterday's health event, and demonstrating that the resulting indicator discriminates well for residual confounding. They later extended it from detection to partial correction (Flanders et al. 2017). Subsequent work has reviewed the framework and developed identification results using negative control exposures and outcomes jointly (Shi et al. 2020), and Hu et al. (2023) provided a recent treatment specific to time-series studies.

The year-shift control used here belongs to this family. It differs from future exposure in one respect that matters for the present purpose, since shifting by whole years preserves the seasonal cycle of the exposure series, so the control targets residual seasonal confounding specifically rather than confounding in general.

2.5 Machine learning and neural approaches

Applications of machine learning in this area divide into two groups. The larger group concerns exposure prediction, forecasting pollutant concentrations from meteorological and temporal covariates. The smaller and more relevant group applies machine learning to the exposure-health relationship itself. Sun et al. (2023) combined machine learning with a distributed lag non-linear model to construct an early-warning model for respiratory mortality under mixed pollutant and meteorological exposure. A recent review of machine learning applied to air pollution and hospital admissions noted the predominance of random forests, neural networks and recurrent networks, and identified capture of lag effects as the central technical problem, with LSTM architectures proposed on the grounds that they model sequential dependence (Abedian Aval et al. 2026). A systematic review of explainable AI in air pollution risk assessment and respiratory outcomes cataloged the growth of this literature between 2020 and 2025 and identified model opacity as the principal barrier to use (Agbehadji and Obagbuwa 2025).

Published evaluations in this group report goodness of fit and, sometimes, held-out prediction. No published example in this application was found reporting a negative control result for the proposed model, which is the gap the comparison in Table 4 addresses.

2.6 Viral seasonality as a competing driver

Respiratory syncytial virus is the leading cause of hospitalization for respiratory disease among children under five years of age, most commonly during the first year of life, and its seasonality in Brazil is well characterized and regionally specific (Freitas and Donalisio 2016). A national analysis using wavelet and Fourier methods identified stable annual cycles with regional differences in timing (Freitas and Donalisio 2016). A temporal-spatial analysis of bronchiolitis hospitalizations across all Brazilian states, drawn from the same hospital system used here, reported the seasonal RSV peak in the city of Rio de Janeiro from March to May, against April/May to July/August in São Paulo and May to July in Salvador (Pereira et al. 2023). National surveillance for 2019 to 2023 placed RSV peaks between March and June (Julian et al. 2026).

This matters for the present design in two ways. It explains the April–May peak in the outcome series, shown in Table 1 and Figure 1, and it establishes that the outcome peak precedes the mid-year PM10 peak. Had the two coincided, separating the exposure association from season would have been considerably harder.

2.7 What this study adds

Existing Brazilian studies establish an association but mostly use short lag windows and small daily counts. The methodological literature on distributed lag models is mature, and the literature on negative controls in time series is well developed, but the two are rarely combined in applied work. Machine learning methods are being proposed for this application without the external validation that the negative-control literature makes available. This study fits fifteen estimators to one dataset with one confounder specification, reports a negative control for everyone, and uses the control to select seasonal adjustment.

3. Data

3.1 Hospital admissions

Records came from SIH-SUS, the billing system of Brazil's public health service, maintained by DATASUS. Monthly reduced-record (RD) files for Rio de Janeiro state were downloaded from the public FTP server, and admissions were retained if they met three conditions, namely municipality of residence Rio de Janeiro (IBGE code 330455), age under five years, and a principal diagnosis in ICD-10 J04, J12–J18, J20–J21, J45 or J46. This set covers acute laryngitis and tracheitis, pneumonia, acute bronchitis and bronchiolitis, and asthma, following earlier Brazilian studies of pediatric respiratory morbidity (Nascimento et al. 2025; Negrisoli and Nascimento 2013; Tuan et al. 2015).

Two features of the source required attention. The files are organized by billing month rather than admission month, so an admission appears in the file for the month the hospital invoiced it, and long stays are invoiced after the month of admission. Restricting downloads to billing months inside the study period would have thinned the final months of the series. Billing months through April 2020 were downloaded and filtered on admission date, recovering 21,329 records dated 2019 that would otherwise have been absent. The series ends on 31 December 2019, before COVID-19 control measures produced a documented fall in pediatric respiratory admissions in Brazil, reported at over 70% for acute bronchiolitis in children under one year (Friedrich et al. 2025).

Age appears twice in these files, once as a number with a companion unit code and once as a date of birth. Age was computed from date of birth and admission date, and the coded fields were used as a cross-check. The two agreed to within one year for 99.85% of records, and age was resolved for every record.

Days with no admissions were retained as zeros. Table 1 shows that only 0.2% of days had none. The daily admissions series is plotted in Figure 1(a).

3.2 Air quality and weather

Hourly measurements came from MonitorAr-Rio, the municipal monitoring network, published on the city open data portal. Eight stations operated over the study period, namely Centro, Copacabana, São Cristóvão, Tijuca, Irajá, Bangu, Campo Grande and Pedra de Guaratiba. PM10 and ozone were measured at all eight, sulfur dioxide and carbon monoxide at most, nitrogen dioxide at four, and PM2.5 at Irajá only. Coverage by pollutant is given in Table 1.

The portal records these data in Coordinated Universal Time for the study period, switching to local time only in November 2023. Every observation was shifted back three hours to Brazilian local time before aggregation to days, reassigning 12.5% of observations to a different calendar day. This is the expected share, since three of every twenty-four hours cross midnight under a three-hour shift.

A station-day mean was computed only where at least 16 valid hourly readings were available. The city-wide series is the unweighted mean of the available station means, and PM10 drew on all eight stations on the median day.

PM10 was available on 99.2% of study days. The 27 missing days fell in two runs, the longer of which lasted 14 days. Because a missing day invalidates the lag window of the following fortnight, 3,218 of 3,287 days (97.9%) carry a complete 14-day exposure history and enter the models. Figure 2 plots the resulting exposure and meteorological series.

3.3 Descriptive statistics

Table 1 summarizes the outcome, exposure and meteorological series. Admissions averaged 11.30 per day and were strongly seasonal, peaking in April and May, consistent with the reported March-to-May RSV season in Rio de Janeiro (Pereira et al. 2023). PM10 averaged 37.79 μg/m³ and peaked in July, two months after the outcome.

The series contained 37,137 admissions over the nine-year window, with only 0.2% of days recording none, and its raw variance-to-mean ratio of 4.47 indicated overdispersion well beyond a Poisson process. The seasonal contrast was wide. Monthly mean admissions rose from 5.5 per day in January (February 5.9, March 10.0) to 19.6 in April and 21.2 in May, then fell away through June 17.6, July 13.6, August 10.1, September 8.9, October 8.1, November 8.0 and December 6.9. Monthly mean PM10 followed a different calendar, moving from 34.5 μg/m³ in January (February 36.7, March 32.1, April 34.0, May 37.9) to 42.7 in June and a July maximum of 46.1, then 42.7 in August, 42.8 in September, 38.3 in October, a November minimum of 31.3 and 34.2 in December. The daily admissions and PM10 series are plotted in Figure 1(a) and Figure 1(b) respectively, and their seasonal profiles are compared in Figure 1(c) and Figure 1(d), where the two-month offset between the April–May outcome peak and the July exposure peak is evident.

Table 1. Descriptive statistics, Rio de Janeiro, 1 January 2011 to 31 December 2019 (3,287 days).

Variable	Coverage	Mean	SD	Min	P25	Median	P75	P99	Max
Admissions, under 5 (per day)	100%	11.30	7.10	0	—	10	—	—	52
PM10 (µg/m³)	99.2%	37.79	14.09	10.79	27.74	35.06	45.37	79.88	108.91
PM2.5 (µg/m³)	79.9%	17.29	10.84	0.21	9.97	14.81	21.58	52.96	104.30
NO_2 (µg/m³)	88.1%	34.40	11.04	6.95	26.65	32.99	40.27	67.80	88.47
O_3 (µg/m³)	99.2%	30.47	11.86	3.10	21.72	29.51	37.87	63.65	84.67
SO_2 (µg/m³)	99.2%	4.50	2.68	0.22	2.58	3.89	5.85	13.29	20.45
CO (ppm)	99.2%	0.34	0.12	0.07	0.26	0.32	0.39	0.76	1.16
Temperature (°C)	99.2%	26.06	3.39	15.91	23.55	25.86	28.50	33.13	35.33
Relative humidity (%)	99.2%	70.70	10.19	39.42	63.72	70.90	77.48	92.67	97.24

Correlations among candidate exposures are given in Table 2. PM10 correlated 0.62 with NO_2, 0.60 with CO and 0.58 with SO_2, consistent with shared traffic and combustion sources; 0.14 with O_3, 0.20 with temperature and −0.37 with relative humidity.

Table 2. Pearson correlations among daily city-wide exposure and meteorological series, Rio de Janeiro, 2011–2019 (n = 3,260 days with PM10 available).

	PM10	NO_2	O_3	SO_2	CO	Temp	RH
PM10	1.00	0.62	0.14	0.58	0.60	0.20	−0.37
NO_2	0.62	1.00	−0.20	0.50	0.66	−0.07	−0.14
O_3	0.14	−0.20	1.00	0.07	−0.08	0.44	−0.39
SO_2	0.58	0.50	0.07	1.00	0.54	0.32	−0.31
CO	0.60	0.66	−0.08	0.54	1.00	0.15	−0.09
Temp	0.20	−0.07	0.44	0.32	0.15	1.00	−0.54
RH	−0.37	−0.14	−0.39	−0.31	−0.09	−0.54	1.00

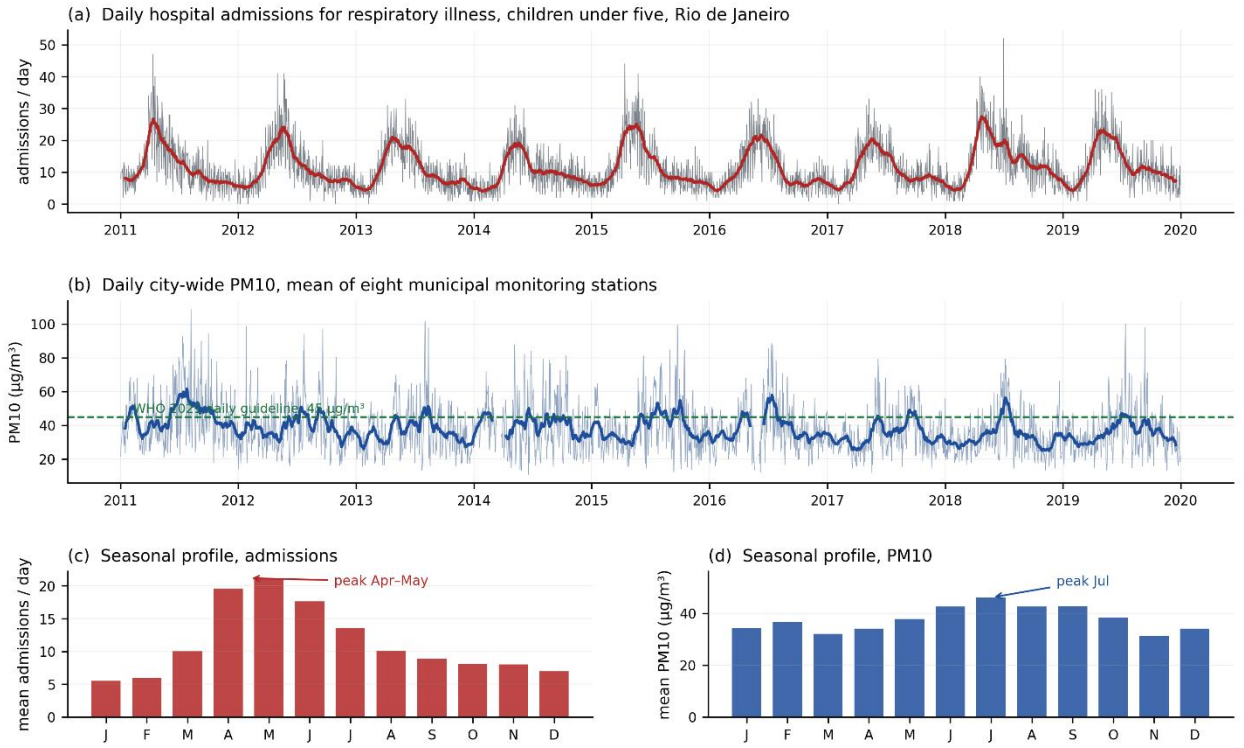


Figure 1. (a) Daily hospital admissions for respiratory illness among children under five in Rio de Janeiro, 2011–2019, with 30-day rolling mean. (b) Daily city-wide PM10, mean of eight municipal monitoring stations, with the WHO 2021 daily guideline

marked. (c) Seasonal profile of admissions, peaking in April and May. (d) Seasonal profile of PM10, peaking in July. The two-month offset between the outcome and exposure peaks assists identification of the exposure association.

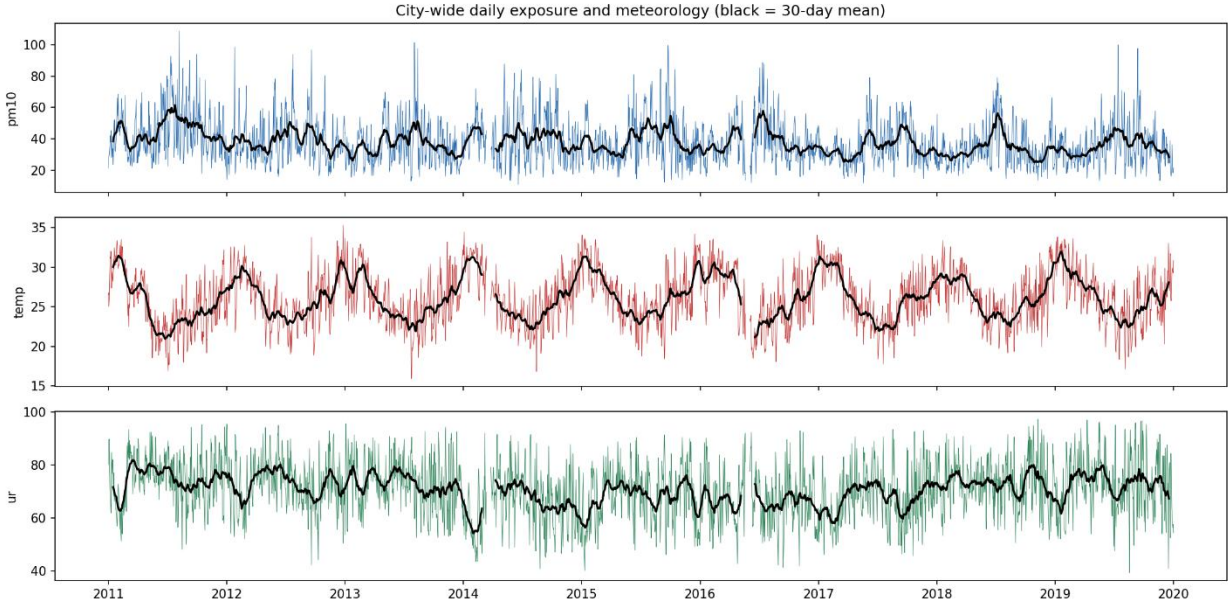


Figure 2. Daily city-wide PM10, temperature and relative humidity, Rio de Janeiro 2011–2019, with 30-day rolling means. Temperature and relative humidity enter the models as confounders, the latter also serving as the omitted covariate in the negative-control calibration of §4.6. Breaks in the PM10 trace correspond to the two gaps described in §3.2.

4. Methods

Figure 3 gives the complete design, covering data sources, processing, the estimator bank, the negative-control evaluation, and the outputs.

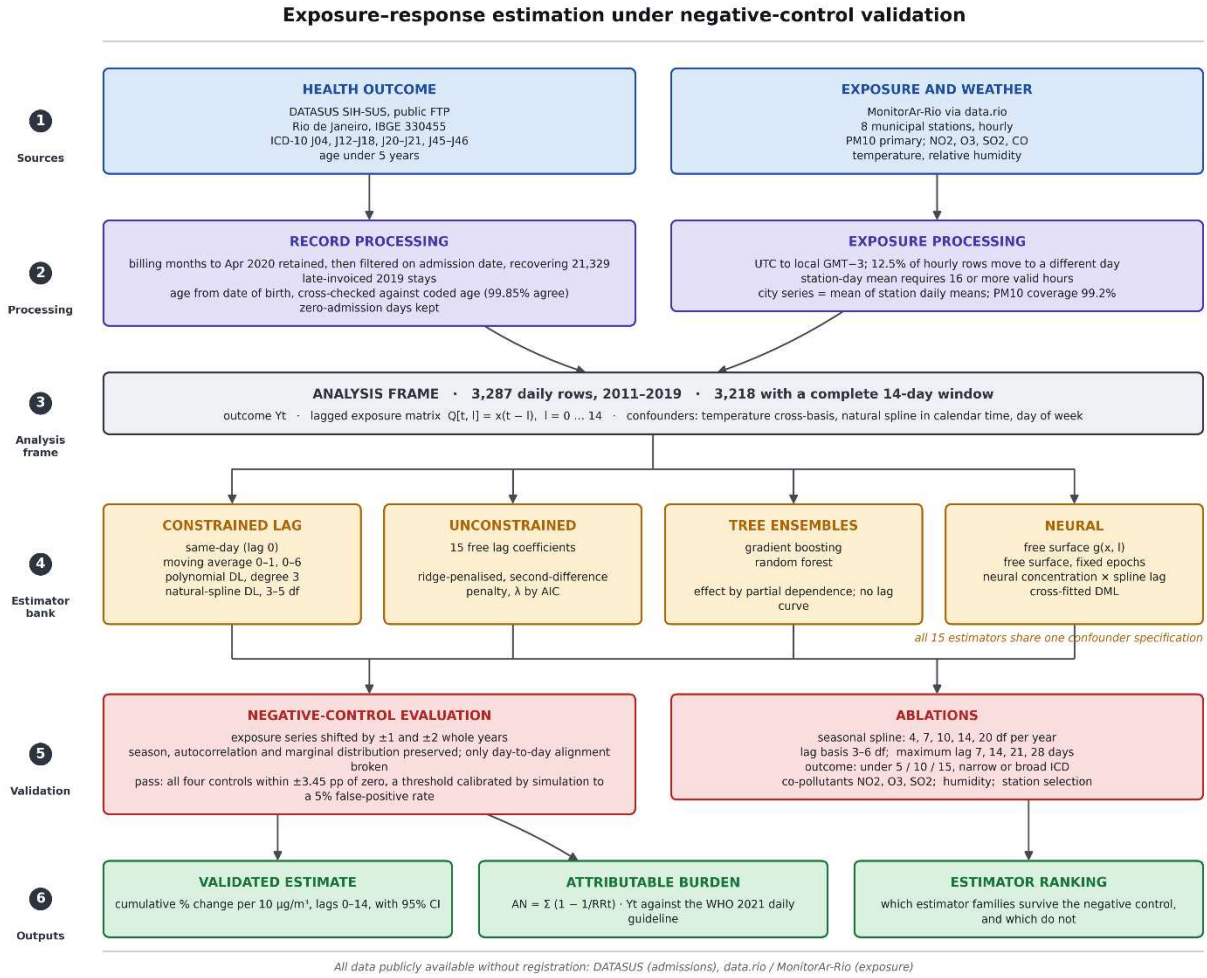


Figure 3. Study design: data sources, record and exposure processing, construction of the analysis frame, the fifteen-estimator bank, negative-control evaluation and ablations, and the three outputs.

4.1 Distributed lag model

Let Y_t be the admission count on day t and x_t the city-wide PM10 concentration. The lagged exposure matrix is defined in Equation (1).

$$Q[t, l] = x_{t-l}, \qquad l = 0,1, \dots, L, \qquad L = 14 \tag{1}$$

Row t of Q holds the fortnight of exposure preceding day t. Following Gasparrini et al. (2010), the exposure-lag-response surface is written as a tensor product of a basis v over exposure value and a basis w over lag, as given in Equation (2).

$$\log \mathrm{RR}\,(x, l) = \sum_{j=1}^{J} \sum_{k=1}^{K} \beta_{jk}\; v_j(x)\, w_k(l) \tag{2}$$

The resulting $J \times K$ cross-basis columns entered the regression. For particulate exposure the basis v was taken as the identity, standard practice for PM, and w as a natural cubic spline with four degrees of freedom, so $J = 1$ and $K = 4$. The lag basis includes a constant so that an effect uniform across lags is representable;

where a spline was used on the concentration axis, as in the temperature term, it was centered so it does not compete with the model intercept.

$L = 14$ was chosen because respiratory admissions following pollution exposure are generally reported within about two weeks (Bhaskaran et al. 2013; Weinmayr et al. 2010), and because the window must be short enough that most days retain a complete exposure history. This choice is examined empirically in §5.3 and Figure 5.

4.2 Regression and adjustment

The descriptive statistics in Table 1 show that the outcome is over dispersed relative to Poisson, so a negative binomial form was assumed, as given in Equation (3).

$$Y_t \sim \mathrm{NB}(\mu_t, \theta), \qquad \mathrm{Var}(Y_t) = \mu_t + \frac{\mu_t^2}{\theta} \tag{3}$$

The mean was modeled as in Equation (4).

$$\log \mu_t = \alpha + \mathrm{cb}\big(\mathrm{PM}_{10,t}; L\big) + \mathrm{cb}(\mathrm{temp}_t; L) + s(t;\ 7\ \mathrm{df/year}) + \gamma \mathrm{DOW}(t) \tag{4}$$

Here $cb(\cdot)$ denotes the cross-basis of Equation (2), $s(\cdot)$ a natural cubic spline in calendar time, and γ day-of-week indicators. The specification followed standard practice for time-series regression in environmental epidemiology (Bhaskaran et al. 2013). The dispersion θ was estimated from a preliminary Poisson fit, as in earlier Brazilian analyses of the same outcome (Nascimento et al. 2025).

Temperature entered through its own non-linear cross-basis over the same window. The calendar-time spline removed seasonal and long-term variation so the exposure estimate rests on short-term day-to-day fluctuation.

Cumulative effects are reported as the percentage change in admissions for a 10 μg/m³ increase in PM10 above a reference x_0, taken as the 25th percentile of the observed distribution (27.74 μg/m³), cumulated over lags 0 to L. Writing $\boldsymbol{z}$ for the corresponding contrast vector in the cross-basis coefficients $\boldsymbol{\beta}$ with covariance $\boldsymbol{V}$, the cumulative effect and its variance are given in Equation (5).

$$\Delta\% = 100[\exp(\boldsymbol{z}^\top \boldsymbol{\beta}) - 1], \qquad \mathrm{Var}(z^\top \beta) = z^\top V z \tag{5}$$

Confidence intervals at the 95% level follow from the delta method applied to Equation (5).

4.3 The year-shift negative control

A negative control is an analysis whose answer is known in advance (Lipsitch et al. 2010). Flanders et al. (2011, 2017) constructed one for air pollution time series from future exposure. A related construction was used here, shifting the exposure series by δ days while leaving admissions in place, and refitting Equation (4) with $x(t + \delta)$ in place of x_t.

The whole-year shift is deliberate. It preserves the seasonal cycle of the exposure series, its autocorrelation and its marginal distribution, destroying only the correspondence between a particular day's pollution and that day's admissions. Because season is preserved, an association surviving the shift is most plausibly residual seasonal confounding, the specific threat in this design.

An estimator passes if all four shifted estimates fall within a threshold of zero. The threshold was calibrated by simulation rather than chosen by convention: as described in §4.6, the worst-case control has a null distribution with mean 1.85 pp and 95th percentile 3.45 pp under the confounder structure of these data, so a threshold of 3.45 pp corresponds to a false-positive rate of 5%. The worst single control is reported rather than the mean, because two large controls of opposite sign can average near zero while indicating an unstable estimator. Results for all estimators are in Table 4 and Figure 7.

The 3.45 percentage point threshold was fixed by the calibration described in §4.6 before any estimator was compared. All four control values for every estimator are reported in Table 4 so that a reader may apply a stricter criterion.

4.4 Estimators compared

Fifteen estimators were fitted, all sharing the identical confounder matrix of Equation (4), so any difference between them comes from the exposure specification alone. They are listed in Table 4.

Seven of the fifteen estimators impose a constrained lag structure. These are same-day exposure and moving averages over lags 0–1 and 0–6, as used in several earlier Brazilian studies (Negrisoli and Nascimento 2013; Tuan et al. 2015), together with a polynomial distributed lag of degree three and natural spline distributed lags with three, four and five degrees of freedom, of which the four-degree specification is the primary model.

Two further estimators leave the lag structure unconstrained. The first fits fifteen free lag coefficients, which is the most flexible lag structure available within a generalized linear model, and the second fits the same model with a second-difference ridge penalty across lags whose strength was selected by AIC.

Two tree ensembles were also fitted. Gradient boosting and random forest were applied to Q of Equation (1) together with the confounders, and the exposure effect was read off by partial dependence. These two estimators represent the tree-based methods common in the applied machine learning literature (Abedian Aval et al. 2026).

Four neural estimators complete the bank. The first replaces the bilinear form of Equation (2) with a multilayer perceptron $g(x, l)$, fitted with roughness penalties in both directions and a one-sided penalty discouraging a decreasing concentration-response. The second constrains the lag dimension to the spline basis w of Equation (2) and learns only the concentration function. The third uses cross-fitted double machine learning, residualizing outcome and exposure against the confounders before estimating the exposure-response. A fourth variant repeats the first with a fixed number of training epochs and no validation-based early stopping, isolating the contribution of that model-selection step.

Neural models were trained by maximizing the negative binomial likelihood implied by Equation (3), with confounder coefficients initialized at values from a generalized linear model fitted to the training rows only. Uncertainty for the first variant came from 30 moving-block bootstrap replicates with 30-day blocks.

4.5 Attributable burden

For each day whose exposure exceeded a counterfactual threshold x_c, the attributable number is given by Equation (6).

$$\mathrm{AN} = \sum_t \left(1 - \frac{1}{\mathrm{RR_t}}\right) Y_t\,, \qquad \mathrm{RR_t} = \exp\left(\sum_{l=0}^{L} \log \mathrm{RR}\,(x_{t-l},\, l)\right) \tag{6}$$

Each day's contribution was evaluated at the cumulative relative risk implied by that day's own exposure history relative to x_c. This is the backward formulation of Gasparrini and Leone (2014). Because Equation (6) works from observed counts, no population denominator is required. Confidence intervals come from 2,000 Monte Carlo draws of $\boldsymbol{\beta}$ from its estimated sampling distribution. Two thresholds were used, the WHO daily PM10 guideline of 45 µg/m³ (World Health Organization 2021) and the 25th percentile of the observed distribution. Results are in Table 6.

4.6 Calibrating and validating the negative control

Showing that an estimator returns a non-null value under shifted exposure is not the same as showing that one returning a null value is unbiased, and only the second justifies a pass criterion. We therefore calibrated the control by simulation.

Simulated outcomes were generated as log μ_t = baseline(t) + f(exposure) + $c \cdot U_t$, where baseline(t) is the fitted seasonal, temperature and day-of-week structure from the observed data, so the simulated series inherits the real confounder geometry; f(exposure) is a distributed lag effect constructed to give a cumulative increase of exactly 3.00% per 10 µg/m³; and U_t is relative humidity, standardized. Humidity is measured on 99.2% of days, correlates −0.37 with PM10, and is omitted from the analysis model, so c controls how much genuine confounding survives adjustment. A constructed smooth seasonal term would have been absorbed by the 63-degree-of-freedom calendar spline and produced no bias; using a real omitted covariate makes the threat realistic rather than artificial. Counts were drawn as negative binomial with θ = 48, matching the observed dispersion.

Sixty replicates were generated at each of six confounder strengths. For each, the primary model was fitted, the four year-shifted controls computed, and the replicate recorded as flagged if its worst-case control exceeded the threshold. A further 60 replicates were generated with no injected confounding, using an independent set of random seeds, to characterize the null distribution of the worst-case control. These are reported separately from the c = 0 level of the power sweep, and the small difference between the two (1.85 pp against 1.74 pp) indicates the Monte Carlo precision available at 60 replicates.

The same framework was used to test the cross-fitted double machine learning estimator, which returned an extreme result on the observed data. Ten replicates were generated with no injected confounding and a known effect of 3.00%, and both the classical and the cross-fitted estimator were asked to recover it.

5. Results

5.1 Main estimate

A 10 µg/m³ increase in PM10 was associated with a 3.75% increase in admissions cumulated over lags 0 to 14 (95% CI 0.63 to 6.97), evaluated by Equation (5). The estimated dispersion parameter was θ = 48.1, corresponding to a conditional variance-to-mean ratio near 1.2 after adjustment, against a raw ratio of 4.47 in Table 1. Almost all of the apparent overdispersion in the unadjusted series was therefore seasonal. The four year-shifted controls gave −0.73, +1.27, −0.79 and −0.56 pp, averaging −0.20 with a worst case of 1.27, as reported in the first block of Table 4.

The cumulative exposure-response curve is shown in Figure 4(a). Relative to the 25th percentile, the cumulative relative risk rose from 0.96 (0.92 to 0.99) at 16 μg/m³ to 1.07 (1.01 to 1.13) at 45 μg/m³ and 1.21 (1.03 to 1.42) at 80 μg/m³.

5.2 Distribution of the association across lags

Table 4 reports that fitting Equation (4) with same-day exposure alone gave 1.67% (0.58 to 2.77), against 3.75% for the full window. A same-day metric therefore recovered under half the cumulative association. The moving average over lags 0–1 gave 1.04% (−0.28 to 2.39) and over lags 0–6 gave 1.50% (−0.56 to 3.60); both intervals included zero.

Table 3 gives the per-lag relative risks, plotted in Figure 4(b). They were close to constant across the window, with a shallow minimum at lag 4 and a second rise near lag 9. No single lag dominated, and the cumulative effect accumulated from small daily contributions.

Table 3. Lag-specific relative risk of under-five respiratory admission for a 10 μg/m³ increase in PM10 above the reference of 27.74 μg/m³, from the primary model of Equation (4).

Lag (days)	RR	95% CI
0	1.0034	0.9965 – 1.0103
1	1.0030	0.9982 – 1.0078
2	1.0026	0.9991 – 1.0061
3	1.0024	0.9992 – 1.0056
4	1.0022	0.9987 – 1.0058
5	1.0023	0.9986 – 1.0060
6	1.0025	0.9991 – 1.0059
7	1.0028	0.9996 – 1.0060
8	1.0030	0.9997 – 1.0064
9	1.0031	0.9996 – 1.0067
10	1.0030	0.9996 – 1.0064
11	1.0026	0.9995 – 1.0058
12	1.0021	0.9987 – 1.0054
13	1.0014	0.9967 – 1.0060
14	1.0006	0.9940 – 1.0073
Cumulative, lags 0–14	**1.0375**	**1.0063 – 1.0697**

These results bear directly on the Brazilian studies that used five- or six-day windows (Nascimento et al. 2025; Negrisoli and Nascimento 2013; Tuan et al. 2015), and are consistent with the observation of Weinmayr et al. (2010) that PM10 associations appeared stronger in studies using longer lags.

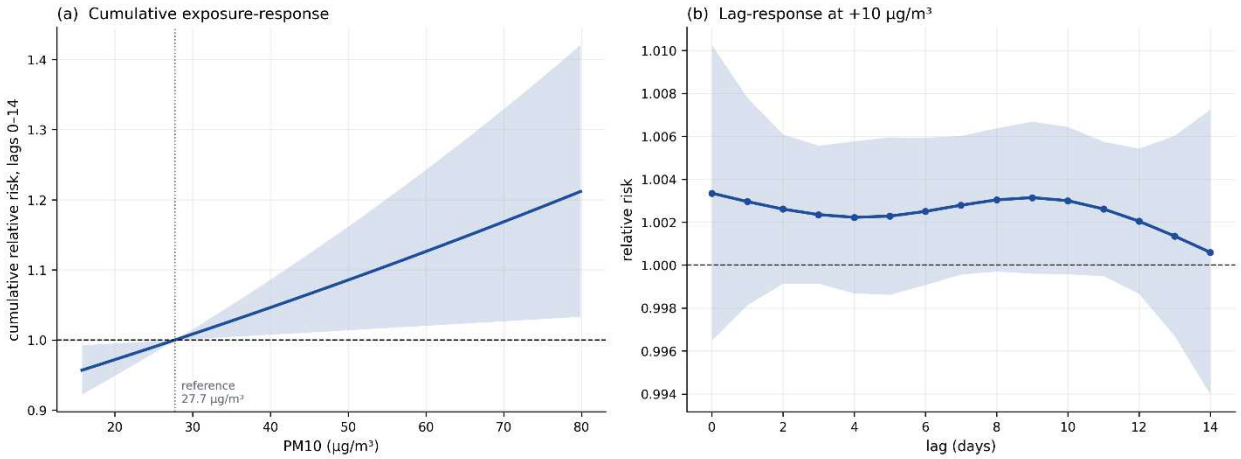


Figure 4. (a) Cumulative exposure-response curve for under-five respiratory admissions against PM10 over lags 0–14, relative to the 25th percentile of 27.74 µg/m³. (b) Lag-response curve for a 10 µg/m³ increase above the same reference. Shaded bands are 95% confidence intervals from the delta method.

5.3 Sensitivity to lag specification

Table 5 shows that four lag bases gave nearly identical answers, 3.75% at three degrees of freedom, 3.75% at four, 3.96% at five and 3.96% at six, with worst-case controls between 1.27 and 1.33 pp throughout. The first three of these are carried into the full estimator comparison in Table 4, alongside the polynomial specification, which gave 3.78%, and the unconstrained fifteen-parameter model, which gave 4.17% (1.02 to 7.41); both passed the control. The ridge penalty selected by AIC was negligible, at $\lambda = 0.01$, and gave 4.21%, indicating the data do not call for shrinkage across lags.

The maximum lag L was varied from 7 to 28 days, and Table 5 reports the result at each window. The estimate rose from 2.65% (0.39 to 4.95) at seven days to 3.75% at fourteen, consistent with an association that accumulates across the window rather than concentrating on any single day. At 21 days the estimate was −0.71% and at 28 days −3.57%, with worst-case controls of 3.22 pp and 4.79 pp. The 28-day specification fails the criterion outright, and in both cases the estimate converged on the mean of its shifted controls, as Figure 5 shows, indicating that the model was no longer separating an exposure association from what the misaligned series produces. A 14-day window is therefore the longest for which the specification remains identified on these data, and the choice of L is determined by the negative control rather than by convention.

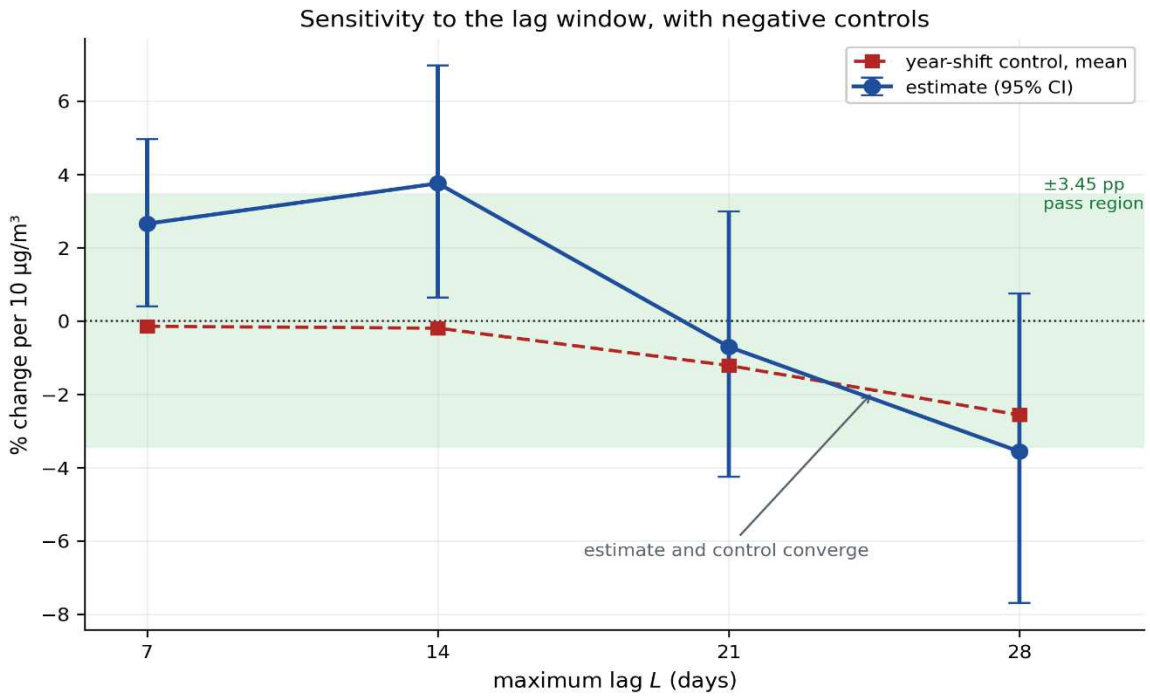


Figure 5. Cumulative effect estimate with 95% confidence interval (solid) and the mean of its four year-shift negative controls (dashed) against the maximum lag *L*. The shaded band marks the ±3.45 pp pass region. Estimate and control converge beyond 14 days, indicating loss of identification rather than absence of an association.

5.4 Sensitivity to seasonal adjustment

Table 5 and Figure 6 show that the estimate rose monotonically with seasonal adjustment, giving −0.84% at four degrees of freedom per year, 3.75% at seven, 5.44% at ten, 7.81% at fourteen and 6.15% at twenty.

The four-degree specification deserves attention. Its point estimate of −0.84% (−3.84 to 2.26) would ordinarily be read as showing no effect. Its year-shifted controls averaged −6.53 pp with a worst case of 7.24, a large association recovered from exposure carrying no signal. The apparent null came from adjustment failing, and nothing in the model's own output revealed it. At twenty degrees of freedom the controls also degraded (mean +2.06, worst case 3.57). That worst-case value sits within the Monte Carlo uncertainty of the 3.45 pp threshold, so the twenty-degree specification is treated as borderline rather than failing. Controls were closest to null at seven, ten and fourteen degrees of freedom, the range over which the estimate was stable and positive, as the two series in Figure 6 show.

AIC preferred ten degrees of freedom per year (17,218.9, against 17,241.8 at seven and 17,221.4 at fourteen). Each specification estimates its own dispersion parameter, rising from $\theta = 48.1$ at seven degrees of freedom to 55.0 at ten and 63.0 at fourteen, so these AIC values are not strictly comparable and are reported as indicative rather than decisive. Seven is reported as primary because it is the field convention (Peng et al. 2006) and the most conservative of the three defensible estimates, with ten and fourteen reported alongside in Table 5. The headline figure should be read as a lower bound.

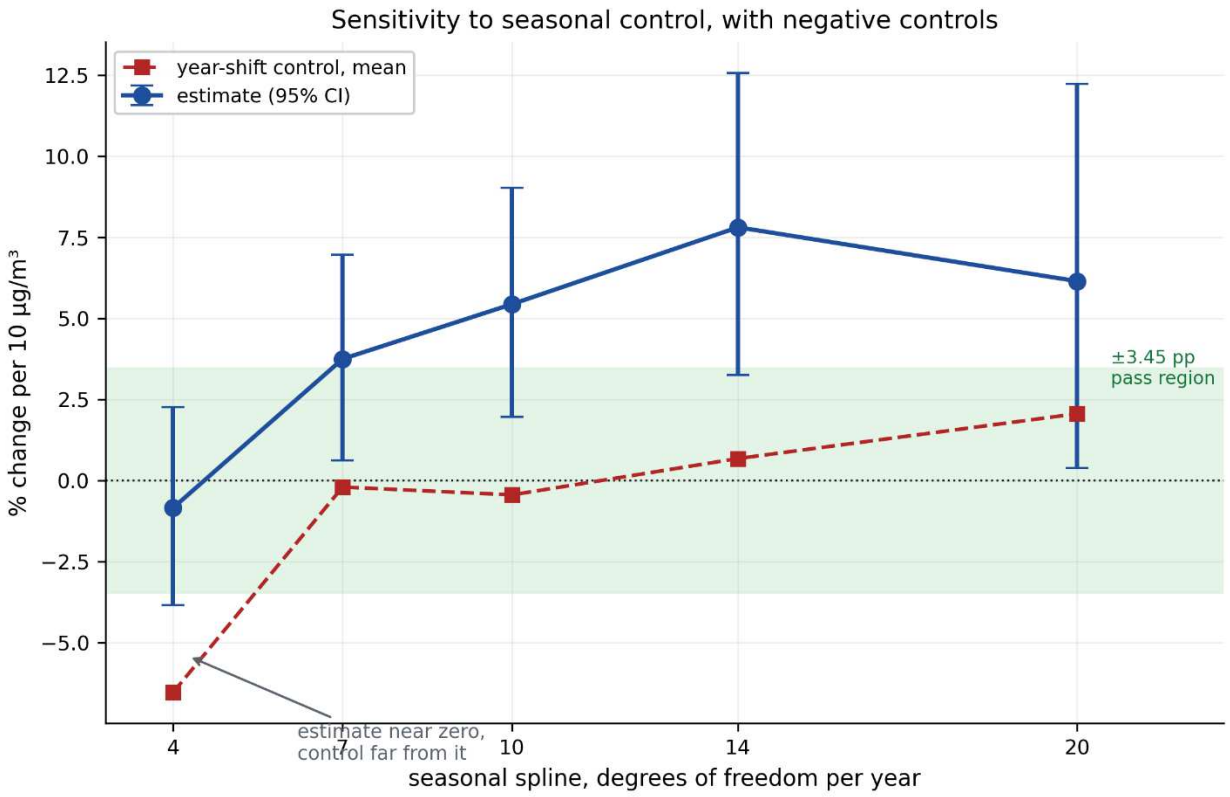


Figure 6. Effect estimate with 95% confidence interval (solid) and mean year-shift negative control (dashed) against seasonal spline degrees of freedom per year. The shaded band marks the ±3.45 pp pass region for the control. At four degrees of freedom the estimate is near zero while the control is not, indicating broken identification rather than absence of an effect.

5.5 Comparison of fifteen estimators

Table 4 gives every estimator with its effect estimate and four shifted controls; Figure 7 plots estimate against worst-case control failure. Eleven of fifteen passed.

Table 4. Effect estimate and year-shift negative controls for fifteen estimators, all fitted to the 3,218 analyzable days with the identical confounder matrix of Equation (4). Shifted columns give the estimate obtained when the exposure series is displaced by the stated interval. An estimator passes if all four shifted estimates lie within ±3.45 pp of zero.

Estimator	Effect (%)	95% CI	−2 y	−1 y	+1 y	+2 y	Worst |shift|	Pass
Same-day (lag 0)	1.67	0.58 – 2.77	−0.26	0.80	−0.05	0.09	0.80	Yes
Moving average 0–1	1.04	−0.28 – 2.39	−0.39	0.86	−0.08	0.38	0.86	Yes
Moving average 0–6	1.50	−0.56 – 3.60	−1.15	0.81	−1.00	0.45	1.15	Yes
Polynomial DL (deg 3)	3.78	0.65 – 7.00	−0.72	1.28	−0.78	−0.57	1.28	Yes
Spline DL, 3 df	3.75	0.63 – 6.97	−0.74	1.29	−0.79	−0.55	1.29	Yes
Spline DL, 4 df (primary)	**3.75**	**0.63 – 6.97**	**−0.73**	**1.27**	**−0.79**	**−0.56**	**1.27**	**Yes**
Spline DL, 5 df	3.96	0.83 – 7.19	−0.56	1.32	−0.66	−0.78	1.32	Yes
Unconstrained DL (15 par.)	4.17	1.02 – 7.41	−0.54	1.36	−0.69	−1.01	1.36	Yes
Ridge DL (λ = 0.01)	4.21	—	−0.55	1.37	−0.70	−0.95	1.37	Yes
Gradient boosting	1.31	—	0.51	−0.06	−0.58	0.61	0.61	Yes
Random forest	0.86	—	0.40	0.01	−0.50	−0.10	0.50	Yes

Neural, free surface	12.30	4.66 – 17.04	13.98	7.42	−5.02	5.57	13.98	No
Neural, free surface, fixed epochs	9.74	—	4.18	3.75	−1.79	0.37	4.18	No
Neural, hybrid	8.42	—	3.85	5.53	−13.77	3.40	13.77	No
Neural, cross-fitted DML	16.03	—	10.32	29.39	31.05	21.46	31.05	No

Confidence intervals are from the delta method for the eight unpenalized generalized linear model specifications, and from 30 moving-block bootstrap replicates with 30-day blocks for the neural free surface. No interval is reported for the ridge specification, the two tree ensembles, or the remaining three neural estimators. The ridge estimate is penalized, so no analytic interval is available; none was computed for the others. The two interval procedures used are not equivalent, and widths should not be compared directly across them.

All nine classical specifications passed, with worst-case controls between 0.80 and 1.37 pp. Gradient boosting and random forest also passed, with worst cases of 0.61 and 0.50, but their point estimates were much smaller. Neither produced a confidence interval or a lag-response curve.

Four neural estimators failed. In each case at least one shifted estimate exceeded the classical model's estimate on correctly aligned data, which is the pattern visible in the upper right of Figure 7.

We investigated one candidate explanation. The three early-stopped neural variants used a validation set to decide when to stop training, a search over many candidate models scored on the same data, which none of the other twelve estimators perform. Refitting the free surface for a fixed number of epochs, with no early stopping, reduced its worst-case control from 13.98 to 4.18 pp and its estimate from 12.30% to 9.74%. Early stopping therefore contributes substantially, but the fixed-epoch fit still exceeds the criterion. We were unable to isolate the remainder and report it as an open question.

The free surface was also tested against a block-permutation null, reordering 30-day blocks of the exposure series. Across 30 permutations the estimator returned a mean of 5.84% (SD 5.83), and four of the 30 permuted values exceeded the value from the correct series. With 30 permutations the resolution of the resulting p-value is 1/31, so this is reported as $p \approx 0.16$ and should be read as indicating no clear separation rather than as a precise significance level.

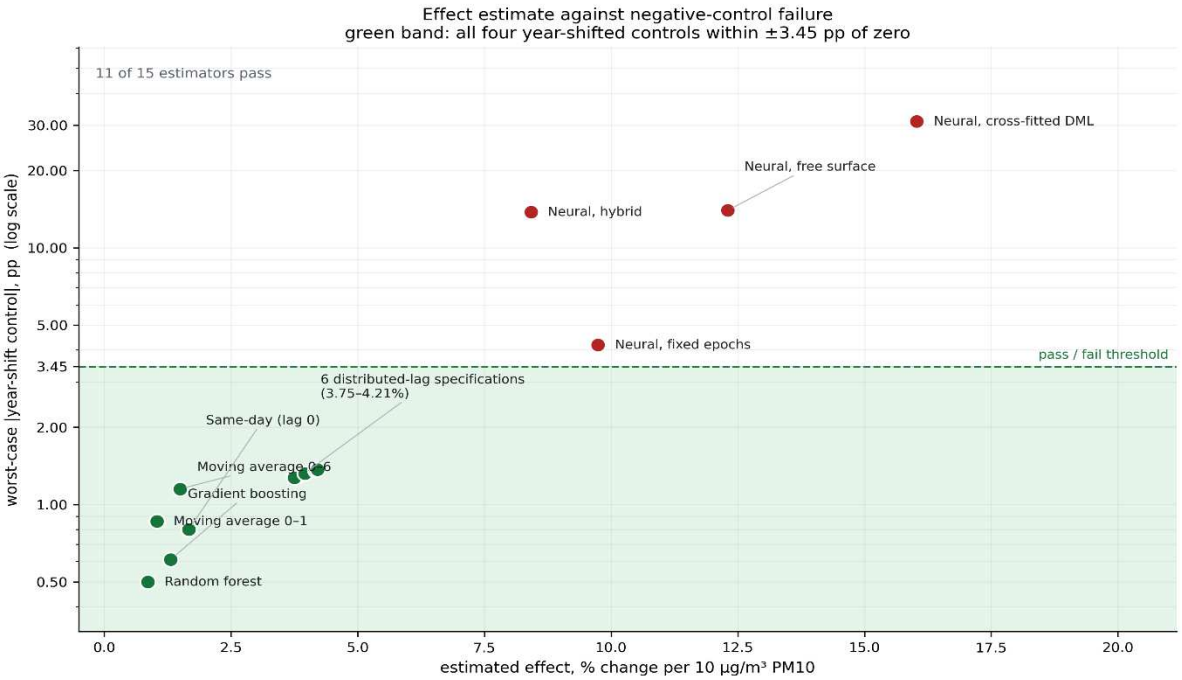


Figure 7. Effect estimate against worst-case absolute year-shift negative control for all fifteen estimators. Green points satisfy the ±3.45 pp criterion; red points do not. The vertical axis is logarithmic. The four failing estimators also report the largest effects.

5.6 Outcome definition and co-pollutants

Table 5 reports all sensitivity analyses, including those requested at review. Widening the diagnostic set to all of ICD-10 Chapter X gave 3.25% (0.28 to 6.31). Widening the age band gave 3.65% (0.68 to 6.71) under ten and 3.03% (0.13 to 6.02) under fifteen. All three remained significant, with worst-case controls of 1.28, 2.24 and 2.48 pp, below the 3.45 pp threshold. The decline with age is consistent with greater susceptibility among younger children (Aithal et al. 2023).

Two alternatives to the eight-station unweighted mean were examined, and Table 5 reports both. Restricting to the four stations in the central zone, on the grounds that Bangu and Campo Grande lie some 40 km away across differing terrain, gave 3.27% (0.14 to 6.50); that series correlates 0.950 with the all-station mean. Weighting each day by the number of contributing monitors, which down-weights the 70 days resting on fewer than four stations, gave 2.85% (1.67 to 4.04), the narrowest interval obtained in any specification. Both passed the negative control, and the estimate is not materially sensitive to how the city-wide exposure series is constructed.

Adjusting for ozone gave 3.58% (0.26 to 7.01) and for sulfur dioxide 2.67% (−1.02 to 6.49). Both models used the same 3,218 days as the primary analysis, since ozone and sulfur dioxide share PM10's daily coverage. Adjustment for nitrogen dioxide was not estimable, since NO_2 is measured at four of eight stations with clustered gaps, and once the affected days were removed the seasonal spline was no longer identified. Ozone is the more informative adjustment in any case, being the co-pollutant least correlated with PM10, at r = 0.14 in Table 2. Relative humidity, available on 99.2% of days and correlated −0.37 with PM10, was added in two forms. As a same-day spline it gave 3.72% (0.59 to 6.94), a change of 0.03 pp from the primary estimate; as a lagged cross-basis matching the treatment of temperature it gave 2.73% (−0.50 to 6.06). Humidity adjustment is therefore immaterial to the estimate, and the simpler specification is retained.

Table 5. Ablations. Each row varies one element of the primary specification of Equation (4), holding the others fixed, and reports the effect estimate with its year-shift controls. Pass is assessed against the 3.45 pp threshold of §5.8. The 20 degrees-of-freedom row is marked borderline because its worst-case control of 3.57 pp lies within the Monte Carlo uncertainty of that threshold, which is estimated from 60 replicates.

Axis	Value	Effect (%)	95% CI	n days	Shift mean	Worst |shift|	Pass
Seasonal df/year	4	−0.84	−3.84 – 2.26	3,218	−6.53	7.24	No
	7 (primary)	**3.75**	**0.63 – 6.97**	3,218	**−0.20**	**1.27**	Yes
	10	5.44	1.97 – 9.03	3,218	−0.44	1.52	Yes
	14	7.81	3.25 – 12.57	3,218	0.68	1.04	Yes
	20	6.15	0.39 – 12.24	3,218	2.06	3.57	Borderline
Lag basis df	3	3.75	0.63 – 6.97	3,218	−0.20	1.29	Yes
	4 (primary)	**3.75**	**0.63 – 6.97**	3,218	**−0.20**	**1.27**	Yes
	5	3.96	0.83 – 7.19	3,218	−0.17	1.32	Yes
	6	3.96	0.83 – 7.20	3,218	−0.17	1.33	Yes
Outcome	**under 5, narrow ICD (primary)**	**3.75**	**0.63 – 6.97**	3,218	**−0.20**	**1.27**	Yes
	under 5, broad ICD	3.25	0.28 – 6.31	3,218	−0.40	1.28	Yes
	under 10, narrow ICD	3.65	0.68 – 6.71	3,218	−0.92	2.24	Yes
	under 15, narrow ICD	3.03	0.13 – 6.02	3,218	−1.04	2.48	Yes
Co-pollutant	+ O_3	3.58	0.26 – 7.01	3,218	−0.22	1.24	Yes
	+ SO_2	2.67	−1.02 – 6.49	3,218	−0.15	1.48	Yes
	+ NO_2	not estimable	—	—	—	—	—
Maximum lag	7 days	2.65	0.39 – 4.95	3,219	−0.15	1.22	Yes
	14 days (primary)	**3.75**	**0.63 – 6.97**	**3,218**	**−0.20**	**1.27**	**Yes**
	21 days	−0.71	−4.26 – 2.98	3,197	−1.22	3.22	No
	28 days	−3.57	−7.70 – 0.74	3,176	−2.57	4.79	No
Station selection	central 4 stations	3.27	0.14 – 6.50	3,218	−0.35	0.72	Yes
	weighted by station count	2.85	1.67 – 4.04	3,218	0.01	1.74	Yes
Confounder	+ humidity, same-day spline	3.72	0.59 – 6.94	3,218	−0.26	1.15	Yes
	+ humidity, lagged cross-basis	2.73	−0.50 – 6.06	3,218	−0.28	1.18	Yes

5.7 Attributable burden

Table 6 gives the attributable burden from Equation (6). PM10 exceeded the WHO daily guideline of 45 µg/m³ (World Health Organization 2021) on 834 of the 3,218 analyzed days, and an estimated 449 admissions (95% CI 80 to 795) were attributable to concentrations above that guideline, 1.24% of admissions analyzed.

Table 6. Attributable burden of under-five respiratory admissions, computed from Equation (6) against two counterfactual exposure thresholds. Total admissions on analyzable days were 36,202.

Counterfactual	Threshold (µg/m³)	Days above	Attributable admissions	95% CI	% of total

WHO 2021 daily guideline	45.00	834	449	80 – 795	1.24
25th percentile of observed	27.74	2,413	1,544	275 – 2,715	4.27

5.8 Calibration and power of the negative control

Under no injected confounding, the worst-case control had mean 1.85 pp, median 1.77 pp, 90th percentile 2.76 pp and 95th percentile 3.45 pp across 60 replicates, as Table 7 reports. The control is therefore not exactly null even when the model is correctly specified, for two reasons that can be quantified: a whole-year shift leaves the shifted series correlated with the true series at r = 0.088 because the seasonal cycle repeats, and PM10 declined over the study period by 0.93 µg/m³ per year, so shifts in opposite directions are not symmetric. Together these produce a small systematic positive control, visible in Table 4 as a −1 y control between +0.80 and +1.37 pp for every classical specification.

Table 7. Calibration and power of the year-shift negative control. At each confounder strength c an omitted covariate was injected into the simulated series and the worst-case control recomputed.

Confounder strength c	Mean worst-case control (pp)	Detection rate	Replicates
0.00	1.74	0.32	60
0.05	1.63	0.25	60
0.10	1.91	0.42	60
0.20	2.63	0.85	60
0.30	3.23	0.88	60
0.45	4.60	1.00	60

Null distribution at c = 0: mean 1.85, median 1.77, 90th percentile 2.76, 95th percentile 3.45 pp. Detection at the recalibrated 3.45 pp threshold. Eighty percent detection at c = 0.188.

The control responded monotonically to injected confounding. Mean worst-case values were 1.74, 1.63, 1.91, 2.63, 3.23 and 4.60 pp at increasing confounder strength, and the detection rate rose from 0.32 to 1.00, passing 80% at c = 0.188. Figure 8 plots both series against confounder strength. Across this range the bias in the effect estimate itself remained between −0.44 and −0.80 pp and was not monotone, so the control detects the presence of an omitted covariate correlated with exposure well before that covariate produces a large bias. It is a test of confounding structure rather than a direct measure of bias, and should be read as such.

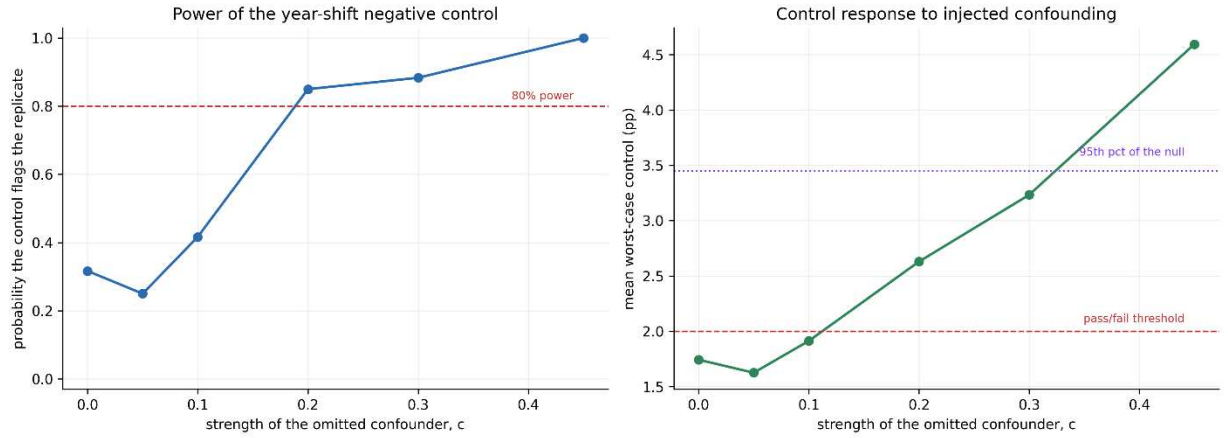


Figure 8. Left: probability that the year-shift negative control flags a replicate, against the strength of an omitted confounder (relative humidity), across 60 replicates per level. Right: mean worst-case control against the same, with the 2 pp pass/fail line used before recalibration and the 95th percentile of the null distribution, 3.45 pp, adopted as the criterion here, both marked.

Recalibrating the threshold to 3.45 pp, the 95th percentile of the null, leaves the estimator comparison unchanged. The eleven passing estimators have worst-case controls between 0.50 and 1.37 pp, below the median of the null distribution; the four neural estimators have 4.18, 13.77, 13.98 and 31.05 pp, beyond its 95th percentile. Neither conclusion is near the boundary.

In the recovery test, the classical model returned 3.46% (SD 1.56) against a true effect of 3.00%, and the cross-fitted estimator returned 4.40% (SD 4.91), with a 95% interval for its mean of 1.36 to 7.44 that contains the truth. Individual replicates ranged from −3.14% to +9.84%. The cross-fitted estimator is therefore approximately unbiased but has a standard deviation roughly three times that of the classical model, and a single estimate is uninformative at this effect size.

6. Discussion

6.1 Main finding

PM10 was associated with hospital admissions for respiratory illness among children under five in Rio de Janeiro over 2011 to 2019, at 3.75% per 10 μg/m³ cumulated over a fortnight, as reported in Table 4. This sits at the upper end of published pooled estimates, which range from 0.4% per 10 μg/m³ for PM10 and pneumonia (Yee et al. 2021) to 1.75% for PM10 and pediatric asthma (Ding et al. 2015). Direct comparison with Gouveia and Fletcher's São Paulo figure of about 9% (Gouveia and Fletcher 2000) is limited by exposure metric, since theirs is a 10th-to-90th percentile contrast, considerably wider than 10 μg/m³.

Being roughly twice the upper end of the pooled range warrants explanation, and four factors plausibly contribute. First, the outcome here is respiratory admission broadly defined, whereas the pooled estimates are specific to pneumonia (Yee et al. 2021) or asthma (Ding et al. 2015; Lim et al. 2016), and a broader outcome captures conditions with differing susceptibility. Second, the analysis is restricted to children under five, the age band in which susceptibility is greatest, while the meta-analyses pool wider age ranges. Third, SIH-SUS records public-sector admissions, which select for households of lower socioeconomic position, among whom baseline respiratory risk and exposure are both higher. Fourth, residual confounding cannot be excluded; the year-shift control constrains the scope for residual seasonal confounding but says nothing about confounding on shorter timescales. Readers should weight the lower bound of the confidence interval, 0.63% per 10 μg/m³, when comparing against the pooled literature.

Because SIH-SUS covers public-sector admissions only, the counts used here exclude children admitted under private insurance, and the burden in Table 6 is a lower bound for the city.

6.2 Timing of the association

The comparison between the same-day model (1.67%) and the full-window model (3.75%) in Table 4 is the practically useful result. Health impact assessments and alert systems keyed to same-day concentration will understate the association by more than half. Table 3 explains why, in that no single day carried the effect, which accumulated in small increments across the fortnight. Several earlier Brazilian analyses used windows of five or six days (Nascimento et al. 2025; Negrisoli and Nascimento 2013; Tuan et al. 2015), and on the present evidence would have recovered part of the association. Weinmayr et al. (2010) reached

a consistent conclusion from a different direction, finding stronger PM10 associations among studies using longer lags.

Table 1 also shows that admissions peak in April and May while PM10 peaks in July, and this matters for the interpretation. The outcome peak coincides with the documented March-to-May RSV season in Rio de Janeiro (Pereira et al. 2023), and had the viral and pollution peaks coincided, separating exposure from season would have been far harder.

6.3 Negative controls as routine practice

Three results argue for reporting negative controls alongside effect estimates, as the methods literature has advocated (Flanders et al. 2011; Flanders et al. 2017; Hu et al. 2023; Lipsitch et al. 2010; Shi et al. 2020).

The first concerns seasonal adjustment. At four degrees of freedom per year the model returned an estimate near zero, which, reported alone, reads as a clean null. Its shifted controls returned −6.53 pp, as Table 5 and Figure 6 show, meaning the specification recovered large associations from exposure carrying no information. Since seasonal degrees of freedom are usually set by convention (Peng et al. 2006), and since practice varies substantially even within single studies (Gouveia et al. 2017), this suggests the control can serve as a selection criterion in its own right, one independent of the model's likelihood.

The second concerns method choice. As Table 4 shows, four neural estimators returned associations two to four times larger than the classical model, which could be read as greater sensitivity. The shifted controls showed otherwise. Held-out likelihood would not have caught this, since across the classical models it varied only between 3.1972 and 3.1983, and the exposure term contributed almost nothing to prediction in any specification, which is expected when the effect is a few percent against strong seasonal variation.

The third concerns the lag window. Estimates at 21 and 28 days were negative and, taken alone, would have been reported as evidence against an association. The 28-day specification fails the control outright, and at both windows the estimate converges on its shifted controls, as Figure 5 shows, identifying loss of identification rather than absence of an effect.

6.4 The neural estimators

This study reports what was observed rather than a general claim about neural methods. Four variants of three architectures were tested, and Table 4 and Figure 7 show that each failed a control which eleven other estimators passed on identical data with identical adjustment. Removing early stopping recovered about 70% of the gap (13.98 → 4.18) for one; the remainder is unexplained. These architectures resemble those in the applied literature (Abedian Aval et al. 2026; Agbehadji and Obagbuwa 2025; Sun et al. 2023), and no published example in this application reporting a negative control was found. Authors proposing such methods are encouraged to report one.

The tree ensembles are a separate case. They passed, but with attenuated estimates and without intervals or a lag curve. They are not drop-in replacements for a distributed lag model, even though they did not produce false positives here.

6.5 Limitations

SIH-SUS covers public-sector admissions only. City-wide averaging of eight monitors ignores within-city exposure gradients, and no individual exposure was assigned, so the design is ecological and estimates

apply to the population rather than to individuals. Table 1 shows that PM2.5 was measured at one station only, so PM10 was the primary exposure and the two cannot be separated. Nitrogen dioxide adjustment was not estimable, as recorded in Table 5. The study covers one city over nine years, and the association may differ elsewhere; the regional heterogeneity in pooled estimates (Lim et al. 2016) suggests caution in generalizing.

The pass criterion is calibrated to a 5% false-positive rate under the confounder structure of these data, using a simulation in which the omitted covariate is a real measured variable (§4.6). The threshold was revised after the comparison, which would be a concern if it altered the conclusions. It does not: no estimator in the comparison of Table 4 lies between the two thresholds, since the largest passing value is 1.37 pp and the smallest failing value 4.18 pp, and the classification is identical under either. Among the ablations of Table 5, one specification, the twenty degrees-of-freedom seasonal spline, falls between them and is marked borderline for that reason. It remains a judgment in two respects. It is absolute rather than scaled to the effect size, and it is applied to four control values per estimator, so with fifteen estimators the sixty control estimates offer some scope for extreme values to arise by chance. Every control value is reported in Table 4 so that readers may apply a different criterion. The control also has a small systematic positive offset, quantified in §5.8, because a whole-year shift does not entirely break the correlation with the true exposure series.

7. Conclusion

PM10 was associated with hospital admissions for respiratory illness in children under five in Rio de Janeiro between 2011 and 2019, at 3.75% per 10 μg/m³ over a two-week window, with more than half the association arising after the day of exposure and around 449 admissions attributable to concentrations above the WHO 2021 daily guideline. Shifting the exposure series by whole years and requiring the resulting estimate to be null distinguished sound specifications from unsound ones in two cases where the models' own fit statistics did not. The first was an under-adjusted seasonal specification whose apparent null was an artifact, and the second was a group of four neural estimators that recovered large associations from misaligned exposure. Reporting such a control alongside the effect estimate costs little and is recommended as routine practice in time-series studies of this kind.

Declarations

Acknowledgements: The authors would like to express their sincere gratitude to the Department of Computer Science and Engineering, Banasthali Vidyapith, Rajasthan, India, for providing the necessary support, guidance, and research environment to carry out this work.

Funding: This is not applicable.

Authors' Contributions: Ankita Mishra: conceived and designed the study, assembled and processed the hospital admission and air quality records, constructed the analysis frame, implemented the distributed lag models and the fifteen-estimator bank, designed and ran the year-shift negative control and the ablation experiments, analyzed the results, prepared the figures and tables, and wrote the original manuscript draft.

Yogesh kumar gupta: contributed to the research methodology, supervised the study, validated the analysis and its findings, reviewed and interpreted the results, critically revised the manuscript for important intellectual content, and provided overall guidance throughout the research. Both authors reviewed, edited, and approved the final version of the manuscript.

Ethical Approval: This is not applicable, as the study used anonymized secondary records from a public administrative database and did not involve human participants, animals, or any clinical trial.

Consent to Participate: This is not applicable.

Consent to Publish: This is not applicable.

Clinical trial number: Not applicable.

Competing Interests: The authors declare no conflict of interest.

Data Availability Statement:
All data used in this study are publicly available and require no registration, institutional agreement or data-use application.
Hospital admission records were obtained from the Sistema de Informações Hospitalares do SUS (SIH-SUS), maintained by DATASUS, Ministério da Saúde, Brazil, accessible at https://datasus.saude.gov.br/transferencia-de-arquivos/ (Início → Serviços → Transferência de Arquivos → Arquivos de Dados → SIHSUS). Monthly reduced-record files for Rio de Janeiro state, named RDRJ⟨yy⟩⟨mm⟩.dbc, were retrieved from the anonymous FTP server at ftp.datasus.gov.br, directory /dissemin/publicos/SIHSUS/200801_/Dados/. Current web browsers no longer support the FTP protocol, so these files must be retrieved with an FTP client or programmatically.
Air quality and meteorological data were obtained from the MonitorAr-Rio network, published by the Prefeitura da Cidade do Rio de Janeiro at https://www.data.rio/datasets/5b1bf5c3e5114564bbf9b7a372b85e17_2. The complete hourly series may be downloaded as CSV from https://opendata.arcgis.com/api/v3/datasets/5b1bf5c3e5114564bbf9b7a372b85e17_2/downloads/data?format=csv&spatialRefId=4326. The underlying feature service is named for the period 2011 to 2018 but the table extends to 2022.
Retrieval and analysis code is available from the corresponding author upon reasonable request.

References

Abedian Aval F, Ataee S, Nemati B, Silva BT, Lopes D, Cirne P, Martins V, Miranda AI, Relvas H (2026) Machine learning on the frontlines of air pollution and public health: revealing the connection with hospital admissions. Atmosphere 17:17. https://doi.org/10.3390/atmos17010017

Agbehadji IE, Obagbuwa IC (2025) Explainable artificial intelligence and machine learning for air pollution risk assessment and respiratory health outcomes: a systematic review. Atmosphere 16:1154. https://doi.org/10.3390/atmos16101154

Aithal SS, Sachdeva I, Kurmi OP (2023) Air quality and respiratory health in children. Breathe 19:230040. https://doi.org/10.1183/20734735.0040-2023

Bhaskaran K, Gasparrini A, Hajat S, Smeeth L, Armstrong B (2013) Time series regression studies in environmental epidemiology. Int J Epidemiol 42:1187-1195. https://doi.org/10.1093/ije/dyt092

de Oliveira Fernandes MA, Andreão WL, Maciel FM, de Almeida Albuquerque TT (2020) Avoiding hospital admissions for respiratory system diseases by complying to the final Brazilian air quality standard: an estimate for Brazilian southeast capitals. Environ Sci Pollut Res 27:35889-35907. https://doi.org/10.1007/s11356-020-07772-x

Ding L, Zhu DJ, Peng DH (2015) Meta-analysis of the relationship between particulate matter (PM10 and PM2.5) and asthma hospital admissions in children. Zhonghua Er Ke Za Zhi 53:129-135 (in Chinese). PMID 25876689

Dominici F, McDermott A, Daniels M, Zeger SL, Samet JM (2005) Revised analyses of the National Morbidity, Mortality, and Air Pollution Study: mortality among residents of 90 cities. J Toxicol Environ Health A 68:1071-1092. https://doi.org/10.1080/15287390590935932

Flanders WD, Klein M, Darrow LA, Strickland MJ, Sarnat SE, Sarnat JA, Waller LA, Winquist A, Tolbert PE (2011) A method for detection of residual confounding in time-series and other observational studies. Epidemiology 22:59-67. https://doi.org/10.1097/EDE.0b013e3181fdcabe

Flanders WD, Strickland MJ, Klein M (2017) A new method for partial correction of residual confounding in time-series and other observational studies. Am J Epidemiol 185:941-949. https://doi.org/10.1093/aje/kwx013

Freitas AR, Donalisio MR (2016) Respiratory syncytial virus seasonality in Brazil: implications for the immunisation policy for at-risk populations. Mem Inst Oswaldo Cruz 111:294-301. https://doi.org/10.1590/0074-02760150341

Friedrich F, Lumertz MS, Petry LM, Pieta MP, Bittencourt LB, Nunes BB, de Castro e Garcia L, Antunes MOB, Scotta MC, Stein RT, Jones MH, Comaru T, Pinto LA (2025) Seasonality of the incidence of bronchiolitis in infants, Brazil, 2016-2022: an interrupted time-series analysis. Rev Paul Pediatr 43:e2023203. https://doi.org/10.1590/1984-0462/2025/43/2023203

Gasparrini A (2011) Distributed lag linear and non-linear models in R: the package dlnm. J Stat Softw 43:1-20. https://doi.org/10.18637/jss.v043.i08

Gasparrini A (2014) Modeling exposure-lag-response associations with distributed lag non-linear models. Stat Med 33:881-899. https://doi.org/10.1002/sim.5963

Gasparrini A, Armstrong B (2013) Reducing and meta-analysing estimates from distributed lag non-linear models. BMC Med Res Methodol 13:1. https://doi.org/10.1186/1471-2288-13-1

Gasparrini A, Leone M (2014) Attributable risk from distributed lag models. BMC Med Res Methodol 14:55. https://doi.org/10.1186/1471-2288-14-55

Gasparrini A, Armstrong B, Kenward MG (2010) Distributed lag non-linear models. Stat Med 29:2224-2234. https://doi.org/10.1002/sim.3940

Gouveia N, Fletcher T (2000) Respiratory diseases in children and outdoor air pollution in São Paulo, Brazil: a time series analysis. Occup Environ Med 57:477-483. https://doi.org/10.1136/oem.57.7.477

Gouveia N, Corrallo FP, Ponce de Leon AC, Junger W, Freitas CU (2017) Air pollution and hospitalizations in the largest Brazilian metropolis. Rev Saude Publica 51:117. https://doi.org/10.11606/S1518-8787.2017051000223

Hu JK, Tchetgen Tchetgen EJ, Dominici F (2023) Using negative controls to adjust for unmeasured confounding bias in time series studies. Nat Rev Methods Primers 3:66. https://doi.org/10.1038/s43586-023-00249-4

Julian GS, Spinardi JR, Flora AP, Vallur SR, Lefebvre d'Hellencourt F, Liang C, Begier E, Berra TZ, Cortez LV, Vitorasso R, Kyaw MH (2026) Seasonal dynamics and burden of respiratory syncytial virus, influenza, and COVID-19 in Brazil: a national surveillance study (2019-2023). J Glob Health 16:04251. https://doi.org/10.7189/jogh.16.04251

Lim H, Kwon HJ, Lim JA, Choi JH, Ha M, Hwang SS, Choi WJ (2016) Short-term effect of fine particulate matter on children's hospital admissions and emergency department visits for asthma: a systematic review and meta-analysis. J Prev Med Public Health 49:205-219. https://doi.org/10.3961/jpmph.16.037

Lipsitch M, Tchetgen Tchetgen E, Cohen T (2010) Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology 21:383-388. https://doi.org/10.1097/EDE.0b013e3181d61eeb

Nascimento LFC, de Almeida ACSR, Pereira JFG, de Castro MT, dos Santos AOR (2025) Air pollution and respiratory diseases in conurbated cities in the state of São Paulo. Rev Paul Pediatr 43:e2024178. https://doi.org/10.1590/1984-0462/2025/43/2024178

Negrisoli J, Nascimento LFC (2013) Atmospheric pollutants and hospital admissions due to pneumonia in children. Rev Paul Pediatr 31:501-506. https://doi.org/10.1590/S0103-05822013000400013

Peng RD, Dominici F, Louis TA (2006) Model choice in time series studies of air pollution and mortality. J R Stat Soc Ser A Stat Soc 169:179-203. https://doi.org/10.1111/j.1467-985X.2006.00410.x

Pereira EQ, dos Santos MLA, Uchimura TT, Menezes E (2023) Temporal-spatial analysis of hospitalizations for bronchiolitis in Brazil: prediction of epidemic regions and periods for immunization against the respiratory syncytial virus. Rev Paul Pediatr 41:e2021304. https://doi.org/10.1590/1984-0462/2023/41/2021304

Shi X, Miao W, Tchetgen Tchetgen E (2020) A selective review of negative control methods in epidemiology. Curr Epidemiol Rep 7:190-202. https://doi.org/10.1007/s40471-020-00243-4

Sun H, Chen S, Li X, Cheng L, Luo Y, Xie L (2023) Prediction and early warning model of mixed exposure to air pollution and meteorological factors on death of respiratory diseases based on machine learning. Environ Sci Pollut Res 30:53754-53766. https://doi.org/10.1007/s11356-023-26017-1

Tuan TS, Venâncio TS, Nascimento LFC (2015) Air pollutants and hospitalization due to pneumonia among children. An ecological time series study. Sao Paulo Med J 133:408-413. https://doi.org/10.1590/1516-3180.2014.00122601

Weinmayr G, Romeo E, De Sario M, Weiland SK, Forastiere F (2010) Short-term effects of PM10 and NO2 on respiratory health among children with asthma or asthma-like symptoms: a systematic review and meta-analysis. Environ Health Perspect 118:449-457. https://doi.org/10.1289/ehp.0900844

World Health Organization (2021) WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. World Health Organization, Geneva. https://iris.who.int/handle/10665/345329

Yee J, Cho YA, Yoo HJ, Yun H, Gwak HS (2021) Short-term exposure to air pollution and hospital admission for pneumonia: a systematic review and meta-analysis. Environ Health 20:6. https://doi.org/10.1186/s12940-020-00687-7